Relevance, prediction and interpretation for linear models with many predictors.

by

Inge S. Helland

University of Oslo

Abstract

In a recent paper Frank and Friedman (1993) give a thorough discussion and comparison of ridge regression, principal component regression and partial least squares regression focusing mainly upon prediction performance. We will develop the theme of that paper further in two directions. First, asymptotical expressions for the prediction error are developed as a tool for such comparisons, extending results of Helland and Almøy (1994). The concept of relevant components for prediction, defined and discussed in Næs and Helland (1993), turns out to be central in these arguments. Secondly, we give a precise formulation on how latent structure can be interpreted from the point of view of principal component regression and partial least squares. Again the concept of relevant components turns out to be central.

Key words: asymptotical comparisons; latent structure; partial least squares; prediction error; principle component regression; relevant components; relevant factors; ridge regression.

1. Introduction.

During the last decade or so, the field of chemometrics has grown into a strong discipline, using many of the traditional methods of statistics, but also developing its own methods and its own jargon. Partly because different languages are used, the communication between the two disciplines has not been as good as it could have been. In particular, chemometrical methods like partial least squares regression (PLS) have often been looked upon as suspect by the statistical community, while the same methods have been used uncritical and without hesitation by many chemometricians. It is therefore welcome that PLS now has been taken up and compared from many different points of view to more known methods like ridge regression (RR) and principal component regression (PCR) in a recent very thorough paper by Frank and Friedman (1993). Their main conclusion from the point of view of prediction error, seems to be that the differences between the methods are relatively small.

Similar results - comparing a somewhat different set of regression methods - has been reached using asymptotical calculations in Helland and Almøy (1994). The point of departure here has been a population model for the joint covariance structure of all the variables involved, and a restriction posed upon this covariance structure which is formulated in terms of the concept of relevant components, a concept that is further discussed in Helland (1990, 1992) and in Næs and Helland (1993). The arguments and the asymptotical calculations are further developed below. Among other things we show that the hypothesis of m relevant components gives a natural situation where partial least squares regression has better asymptotical prediction performance than ridge regression.

In these investigations the different regerssion methods are only compared from the point of view of prediction performance, which of course is the usual statistical criterion. In the discussion of Frank and Friedman (1993), Svante Wold claims that the analysis of a possible latent structure implied by the data is often more important than the prediction aspect. Partial least squares has been used for this purpose for some years now; see for

instance the monograph by Martens and Næs (1989). The discussion of this is usually very informal and imprecise, however, which may be one reason why many statisticians are suspicious.

One purpose of this paper is to show how a latent structure model expressed in terms of partial least squares coefficients can be related to an ordinary latent structure model in the way a statistician would have formulated it. A concept of relevant factors - closely related to relevant components - then arises in a natural way. But first we concentrate on prediction performance.

2. Prediction by linear combinations.

Consider the usual regression model $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$, where $\varepsilon \sim N(0,\sigma^2\boldsymbol{I})$, and where β is a p-vector of unknown parameters. In addition, assume that $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ has n rows which are independently and identically distributed as $\boldsymbol{x} \sim N(\boldsymbol{0}, \Sigma_{xx})$. (For ease of exposition we assume in this paper that all variables are centered; the general case can easily be developed along the lines of Helland and Almøy (1994). It is also assumed that Σ_{xx} is non-singular.) Let y_0, $\boldsymbol{x}_0$ satisfy the same model, i.e., $y_0 = \boldsymbol{x}_0'\beta + \varepsilon_0$ with $\varepsilon_0 \sim N(0,\sigma^2)$ and also $\boldsymbol{x}_0 \sim N(0, \Sigma_{xx})$. In this setting our task is to predict the new variables y_0 from the new predictors $\boldsymbol{x}_0$ in such a way that the expected prediction error $\mathrm{E}[(y_0 - \hat{y}_0)^2]$ is minimized. When p is relatively large, say of the same order as n, it is well known that ordinary least squares prediction can be considerably improved. The way we will seek to improve it here is by basing the least squares procedure on a smaller number m of linear combinations $\hat{\boldsymbol{R}}'\boldsymbol{x}$, where $\hat{\boldsymbol{R}}$ is an pxm matrix depending upon $\boldsymbol{X}$ and $\boldsymbol{y}$. The asymptotical expected prediction error up to order $1/n$ has been given in Helland & Almøy (1995) for several methods of this class, but only under a special model assumption (m relevant components for prediction; see

below for definition and results). Under general conditions there will also be a zero order contribution to this asymptotical prediction error, which is easily found explicitly.

Theorem 2.1

Let $\hat{y}_0$ be the least squares predictor of y_0 based upon $\boldsymbol{X}\hat{\boldsymbol{R}}$. Assume that $\hat{\boldsymbol{R}} \to_P \boldsymbol{R}$ as $n \to \infty$. Then

$$\mathrm{E}[(y_0 - \hat{y}_0)^2] \to \sigma^2 + \beta' [\Sigma_{xx} - \Sigma_{xx}\boldsymbol{R}(\boldsymbol{R}' \Sigma_{xx}\boldsymbol{R})^{-1}\boldsymbol{R}' \Sigma_{xx}]\beta. \qquad (2.1)$$

Proof.

We have $\hat{y}_0 = \hat{\beta}' \boldsymbol{x}_0$, where $\hat{\beta} = \hat{\boldsymbol{R}}(\hat{\boldsymbol{R}}' \boldsymbol{X}' \boldsymbol{X}\hat{\boldsymbol{R}})^{-1}\hat{\boldsymbol{R}}' \boldsymbol{X}' \boldsymbol{y}$, and

$$\mathrm{E}[(y_0 - \hat{y}_0)^2] = \sigma^2 + \mathrm{E}[(\hat{\beta} - \beta)' \Sigma_{xx}(\hat{\beta} - \beta)] \qquad (2.2)$$

by taking expectation over $\boldsymbol{x}_0$ and y_0. By Slutsky's theorem,

$$\hat{\beta} \to \boldsymbol{R}(\boldsymbol{R}' \Sigma_{xx}\boldsymbol{R})^{-1}\boldsymbol{R}' \sigma_{xy} = \boldsymbol{R}(\boldsymbol{R}' \Sigma_{xx}\boldsymbol{R})^{-1}\boldsymbol{R}' \Sigma_{xx}\beta.$$

Using dominated convergence (a closer estimate of the difference between $\hat{\beta}$ and its limit will be given below) and rearranging terms, equation (2.1) results.

◊

It is obviously desirable to choose $\hat{\boldsymbol{R}}$ and thereby $\boldsymbol{R}$ in such a way that the last term on the righthand side of (2.1) vanishes. This turns out to lead to the weak relevance requirement of Næs & Helland (1993), namely that β belongs to span($\boldsymbol{R}$). However, since this choice of $\boldsymbol{R}$ in general will depend upon unknown parameters, one should also aim at finding it in such a way that deviations from this model condition, measured in terms of deviations in the regression parameter vector, increases the limiting prediction error by an amount which is as

small as possible. This leads to a special case of the strong relevance requirement of Næs & Helland (1993).

Theorem 2.2

Let $f(\beta,\boldsymbol{R}) = \beta'[\Sigma_{xx} - \Sigma_{xx}\boldsymbol{R}(\boldsymbol{R}'\Sigma_{xx}\boldsymbol{R})^{-1}\boldsymbol{R}'\Sigma_{xx}]\beta$ be the last term on the righthand side of equation (2.1).

a) One has $f(\beta,\boldsymbol{R}) = 0$ if and only if $\beta \in \text{span}(\boldsymbol{R})$.

b) For each fixed $d>0$, $\max_{\|\Delta\beta\|=d,\,\beta\in\text{span}(\boldsymbol{R})} f(\beta+\Delta\beta,\boldsymbol{R})$ gets its minimal value over all pxm matrices $\boldsymbol{R}$ if and only if span($\boldsymbol{R}$) also is spanned by m eigenvectors of Σ_{xx} corresponding to the m largest eigenvalues of Σ_{xx} (counting multiple eigenvalues by their multiplicities).

Proof.

Let $\boldsymbol{C} = (\Sigma_{xx})^{\frac{1}{2}}\boldsymbol{R}$ and $\gamma = (\Sigma_{xx})^{\frac{1}{2}}\beta$. Then $f(\beta,\boldsymbol{R}) = \gamma'(\boldsymbol{I}-\boldsymbol{P}_C)\gamma$ with $\boldsymbol{P}_C = \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}$, and it is obvious that $f(\beta,\boldsymbol{R}) = 0$ if and only if γ belongs to span($\boldsymbol{C}$), i.e., $\beta \in \text{span}(\boldsymbol{R})$.

When $\beta \in \text{span}(\boldsymbol{R})$, we have $f(\beta+\Delta\beta,\boldsymbol{R}) = \Delta\beta'(\Sigma_{xx})^{\frac{1}{2}}(\boldsymbol{I}-\boldsymbol{P}_C)(\Sigma_{xx})^{\frac{1}{2}}\Delta\beta$. The maximum over this expression for $\|\Delta\beta\| = d$ is d^2 times the largest eigenvalue of the matrix $(\Sigma_{xx})^{\frac{1}{2}}(\boldsymbol{I}-\boldsymbol{P}_C)(\Sigma_{xx})^{\frac{1}{2}}$. This is minimized if and only if $(\boldsymbol{I}-\boldsymbol{P}_C)$ projects away the eigenvectors corresponding to the largest eigenvalues of Σ_{xx}. Since in this case span($\boldsymbol{R}$)=span($\boldsymbol{C}$), assertion b) follows.

◊

In Næs & Helland (1993) the random vector $z=\boldsymbol{R}'\boldsymbol{x}$ was defined to be weakly relevant for predicting y if the best linear predictor of y given z is equal to the best linear predictor of y given $\boldsymbol{x}$. This was shown to be equivalent to the requirement $\beta \in \text{span}(\boldsymbol{R})$, i.e., Theorem 2.2a).

The vector z was called strongly relevant if a representation of the form $\boldsymbol{x}=\boldsymbol{R}z+\boldsymbol{U}\boldsymbol{v}$ can be found such that $\boldsymbol{R}'\boldsymbol{U}=0$, $\mathrm{cov}(z,\boldsymbol{v})=0$ and $\mathrm{cov}(\boldsymbol{v},y)=0$. This is equivalent to weak relevance plus the requirement that span($\boldsymbol{R}$) should be spanned by eigenvectors of Σ_{xx}. This is an important hypothesis that we shall use several times below.

Hypothesis $\mathrm{H_m}$ **(strong relevance).**

Assume that there are m eigenvectors of Σ_{xx} (all corresponding to different eigenvalues) $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m$, and m nonzero parameters $\eta_1, \eta_2, \ldots, \eta_m$ such that

$$\beta = \sum_{k=1}^{m} \eta_k \boldsymbol{e}_k \tag{2.3}$$

This is close to the condition of Theorem 2.2b). The difference is that in Theorem 2.2 we have an additional inequality constraint: The eigenvectors in span($\boldsymbol{R}$) should correspond to the *largest* eigenvalues of Σ_{xx}. Also, since the requirement given in Theorem 2.2 was uniform over $\Delta\beta$, all eigenvectors with large eigenvalues are needed, also in the case of multiplicities. In (2.3) only one vector β is modelled, hence by rotation it is enough with one eigenvector per eigenvalue.

The strong relevant condition without the inequality constraint was shown in Helland (1990) to be closely connected to the partial least squares algorithm from chemometrics, in fact it was the condition that was obtained from this algorithm by formally letting the number of observations tend to infinity and then asking when the algorithm stops in a natural way after m steps. Maximum likelihood prediction under $\mathrm{H_m}$ was developed in Helland (1992). Asymptotical expressions for the prediction error under $\mathrm{H_m}$ was found for several prediction methods in Helland and Almøy (1994). For the present case with centered x-variables these are:

For PCR: $$E[(y_0-\hat{y}_0)^2]\sim\sigma^2(1+\frac{m}{n})+\frac{1}{n}\sum_{k=m+1}^{p}\lambda_k{}^2\sum_{r=1}^{m}\frac{\eta_r{}^2\lambda_r}{(\lambda_r-\lambda_k)^2}+o(\frac{1}{n}) \tag{2.4}$$

For PLS: $$E[(y_0-\hat{y}_0)^2]\sim\sigma^2(1+\frac{m}{n})+\frac{1}{n}\sum_{k=m+1}^{p}\lambda_k{}^2\left\{\omega_k{}^2\sum_{r=1}^{m}\frac{\eta_r{}^2\lambda_r}{(\lambda_r-\lambda_k)^2}+\sigma^2\left(\frac{1-\omega_k}{\lambda_k}\right)^2\right\}+o(\frac{1}{n}),$$

where $$\omega_k=\prod_{i=1}^{m}(1-\lambda_i{}^{-1}\lambda_k). \tag{2.5}$$

Note that the asymptotical formula for PCR is also valid when the model with m relevant components is not valid if we in addition include the constant term due to departure from the model (see Theorem 2.1).

For PLS the corresponding terms are less transparent, but we have the following result: Let $\boldsymbol{R}=\left[\Sigma_{xx}\beta,\Sigma_{xx}{}^2\beta,\ldots,\Sigma_{xx}{}^m\beta\right]$. Then in general (2.1) holds with this $\boldsymbol{R}$ when the prediction method is PLS with m components. (Use Helland, 1990).

The fact that the asymptotical optimality criterion of Theorem 2.2b) involves an explicit inequality requirement, may seem to give an argument for using principal component regression (sorted after the size of the eigenvalue of $\boldsymbol{X'X}$) rather than partial least squares regression. However, the picture is not quite so simple. From the asymptotical expressions (2.4)-(2.5) it is easy to see that partial least squares regression in many cases does much better than principal component regression also in cases where all the relevant eigenvalues are larger than the irrelevant eigenvalues. This is seen from both figures of that paper and also from Figure 1 below. Of course, these differences in prediction error under the (relevant) model should be balanced against the effect of deviations from this model, quantified as in the proof of Theorem 2.2b).

It is clear that by increasing the number m of columns of $\boldsymbol{R}$, the last term on the righthand side of equation (2.1) cannot increase. Thus whichever model assumption one makes and whatever prediction method one uses, this n-independent contribution to the prediction error will decrease as the number of component included is increased. The

estimation errors (2.4)-(2.5) will typically increase with the number of components. The optimal number, found by crossvalidation or in other ways, aims at minimizing the sum of these contributions.

Specifically, when $\boldsymbol{R}$ is spanned by eigenvectors of Σ_{xx}, we get

$$f(\beta,\boldsymbol{R}) = \sum_{k=m+1}^{p} \lambda_k (\beta' \boldsymbol{e}_k)^2 , \qquad (2.4)$$

where $\boldsymbol{e}_k$ are the eigenvectors and λ_k are the eigenvalues of Σ_{xx}. If the λ_k's in (2.4) are the smallest eigenvalues, or if the projections of β on the corresponding eigenvectors are small, the resulting $f(\beta,\boldsymbol{R})$ will be quite small, justifying that it makes sense to balance it against estimation errors of the order $1/n$.

In Theorem 2.2b) we used a minimax condition for the prediction error as a function of the model deviations $\Delta\beta$. Using essentially the same argument, we get the same solution using a Bayesian condition with a prior $\Delta\beta \sim N(0, c(\Sigma_{xx})^{\kappa})$ for some $c>0$ and $\kappa>-1$ (e.g., $\kappa=0$). Note that neither this prior nor the minimax constraint $\|\Delta\beta\| = d$ are invariant under arbitrary scale changes of the components of β. Hence the whole discussion really assumes that these individual components are on the same or comparable scales. This is a well-known problem both in the literature on principal component regression, on shrinkage methods and on ridge regression. All these methods are invariant under rotations, but not under scale changes.

One argument that has been put forward against the kind of asymptotical calculations given above, is the following: When both the number n of objects and the number p of variables are large, it does not make sense to let only n tend to infinity. As a counterargument against this statement we show that it is possible to find an error bound in the asymptotical calculations involved in Theorem 2.1 which is independent of the size of p. It is an open problem to find similar bounds up to order o(1/n) for concrete prediction methods and expansions of the type (2.4)-(2.5), but the result below - and the fact that the asymptotical calculations agree fairly well qualitatively with simulation results - strongly suggest that this should be possible.

Theorem 2.3

We have $\left|\mathrm{E}[(\hat{\beta}-\beta)'\,\Sigma_{xx}(\hat{\beta}-\beta)]-f(\beta,\boldsymbol{R})\right| \le 2\sqrt{\varepsilon_n f(\beta,\boldsymbol{R})}+\varepsilon_n$, where

$$\varepsilon_n = 2\sigma^2 m/(n-m-1)+4\mathrm{E}\left\{\left\|\Sigma_{xx}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\|\|\boldsymbol{y}\|^2 \min\{2,\delta_n/\left[1-(\sqrt{2}+1)\sqrt{\delta_n}\right]_+\right\}$$

with $\delta_n = \kappa(\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{R})\left\|(\hat{\boldsymbol{R}}-\boldsymbol{R})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{R}}-\boldsymbol{R})\right\|/\left\|\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{R}\right\|$. Here $\kappa(\,)$ denotes condition number (the ratio between the largest and the smallest eigenvalue).

The proof of Theorem 2.3 is given in the Appendix. Note that $\left\|\Sigma_{xx}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\|\|\boldsymbol{y}\|^2$ approaches the constant $\sigma_y^{\,2}$, independently of any regression assumption, as n tends to infinity, and that the compound matrices in the definition of δ_n all have dimension mxm.

Note that δ_n will typically be of order $1/n$, and then (from equation (2.2)) Theorem 2.3 gives an error bound for the expected squared prediction error which is of order $1/n$, when β belongs to span ($\boldsymbol{R}$), otherwise of order $1/\sqrt{n}$.

3. Comparison with ridge regression.

In their simulations, Frank and Friedman (1993) got ridge regression (Hoerl and Kennard, 1970) on the top most of the time with PLS as a good number 2. The regression vector for ridge regression is given by

$$\hat{\beta}_{RR} = (\boldsymbol{X}'\boldsymbol{X}+\nu\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad (3.1)$$

where ν is the ridge parameter. Straightforward calculation using equation (2.2) then gives

$$\begin{aligned} \mathrm{E}[(\hat{y}_0 - y_0)^2] = \sigma^2 + \sigma^2 \mathrm{tr}\mathrm{E}\{\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \nu\boldsymbol{I})^{-1}\Sigma_{xx}(\boldsymbol{X}'\boldsymbol{X} + \nu\boldsymbol{I})^{-1}\} \\ + \nu^2 \mathrm{E}\{\beta'(\boldsymbol{X}'\boldsymbol{X} + \nu\boldsymbol{I})^{-1}\Sigma_{xx}(\boldsymbol{X}'\boldsymbol{X} + \nu\boldsymbol{I})^{-1}\beta\} \end{aligned} . \quad (3.2)$$

Letting n tend to infinity in this formula, and assuming that ν is some unknown function of n then implies

Proposition 3.1.

For ridge regression

$$\mathrm{E}[(\hat{y}_0 - y_0)^2] \sim \sigma^2 + \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{\lambda_k{}^2}{(\lambda_k + n^{-1}\nu)^2} + \left(\frac{\nu}{n}\right)^2 \sum_{k=1}^{p}\frac{\lambda_k \eta_k{}^2}{(\lambda_k + n^{-1}\nu)^2}. \quad (3.3)$$

The function of the term $n^{-1}\nu$ in the denominators is to neutralize small eigenvalues λ_k. For this to be possible it is necessary that ν at least is some multiplum of n. This multiplum cannot be large, however, since otherwise the last sum will be appreciable. Qualitatively, this last sum corresponds to the contribution from deviations from the model by the other class of methods, while the first sum corresponds to the model contributions (2.5)-(2.6). It would be interesting to do systematic comparisons over a range of models of the minimum over m of these model errors plus effects of deviations on the one side and the minimum over ν of the righthand side of (3.3) on the other side. This seems to require extensive numerical calculations, however, and is beyond the scope of the present paper. We content ourselves here by showing the relationship for a particular, but important case by giving the following simple result.

Corollary 3.1.

Suppose that the hypothesis $\mathrm{H_m}$ holds for some $m<p$ and assume that ν/n approaches some limit as $n \to \infty$. Then the expected squared prediction error $E\left[(\hat{y}_0 - y_0)^2\right]$ is asymptotically larger for ridge regression than for each of the methods principal

component regression (m components), and partial least squares regression (m components).

Proof.

If $v/n \to 0$, then the righthand side of (3.3) is asymptotically $\sigma^2(1+p/n)$, the same as ordinary least square and clearly dominating the other methods. If $v/n \to a > 0$, the last sum in (3.3) will give a positive constant contribution in addition to σ^2, a contribution that does not accur in (2.5)-(2.6).

◊

4. Regression methods and latent variables; an example.

One of Svante Wold's points in the discussion of Frank and Friedman (1993), where he tried to promote the PLS-type methods over ridge regression, was that by the former methods one can also construct latent variables describing the variables in much the same way as is done in factor analysis. In fact such latent variable constructions are in very common use in chemometrics; we shall briefly explore the connection to factor analysis in this section and in the next one. As a point of departure we look closer at an example given by Frank and Friedman (1993) in their reply to the discussion by S. Wold, where they make some simulations from the following latent variable model with one dependent variable and p=50 predictor variables:

$$\begin{aligned} y &= \sum_{k=1}^{5} l_k^* + \varepsilon \\ x_j &= l_k^* + \delta_j \qquad (10k-9 \le j \le 10k;\ k = 1,\ldots,5) \end{aligned} \tag{4.1}$$

Here $l_1^*,\ldots,l_5^*,\varepsilon,\delta_1,\ldots,\delta_{50}$ are independent normal variables with $\mathrm{Var}(l_k^*) = 1$ and $\mathrm{Var}(\delta_1) = \ldots = \mathrm{Var}(\delta_{50})$.

From a regression point of view it is not too essential that the model (4.1) contains 5 latent variables $l_1^*,...,l_5^*$. The essential information is that the model contains just one relevant latent variable. Namely, by an orthogonal transformation we find the following equivalent representation:

$$\begin{aligned}
y &= 5u_1^* + \varepsilon \\
x_j &= u_1^* + 4u_2^* + \delta_j \qquad (1 \le j \le 10) \\
x_j &= u_1^* - u_2^* + u_3^* + u_4^* + \delta_j \qquad (11 \le j \le 20) \\
x_j &= u_1^* - u_2^* + u_3^* - u_4^* + \delta_j \qquad (21 \le j \le 30) \\
x_j &= u_1^* - u_2^* - u_3^* + u_5^* + \delta_j \qquad (31 \le j \le 40) \\
x_j &= u_1^* - u_2^* - u_3^* - u_5^* + \delta_j \qquad (41 \le j \le 50)
\end{aligned} \qquad (4.2)$$

where $u_1^*, \ldots, u_5^*$ are independent normal variables with $\mathrm{Var}(u_1^*) = 0.2$, $\mathrm{Var}(u_2^*) = 0.05$, $\mathrm{Var}(u_3^*) = 0.25$ and $\mathrm{Var}(u_4^*) = \mathrm{Var}(u_5^*) = 0.50$. The orthogonal transformation from $u_1^*, \ldots, u_5^*$ to $l_1^*,...,l_5^*$ is obvious from (4.1), and the inverse transformation is also easily found. The covariance structure described by (4.1) gives many coinciding eigenvalues in the x-covariance matrix, and this makes the rotation in the factor space possible. The main point is: With this rotation the coupling between x- and y- variables takes place through just one latent factor, not five as in (4.1).

An effective prediction method should be able to take advantage of such a special structure, and ideally use just 1 component in the prediction. It follows from the results of Helland (1990) that the population version of partial least squares regression does just this. Frank and Friedman (1993) found (with n=40 and crossvalidation) that sample PLS chose 3 components on the average, while PCR chose 5. This is consistent with other simulations that I have seen: In similar situations all methods tend to overestimate the number of components, PLSR less than PCR and related methods.

5. Relevant components and relevant factors.

Factor analysis is a well known statistical method, and there is also a literature connecting factor analysis and regression analysis; se for instance Lawley and Maxwell (1973). It is of interest to study the relationship between these areas and the models and methods of the present paper, in particular the model with m relevant components. The example of the previous section shows that there ought to be some connection, and this is further explored here.

In general assume that $(y, x_1, \ldots, x_p)$ is multinormal with zero expectation and with $\sigma_y^2 = \mathrm{Var}(y)$, $\Sigma_{xx} = \mathrm{V}(x_1, \ldots, x_p)$ and $\sigma_{xy} = \mathrm{Cov}(y, (x_1, \ldots, x_p)')$. We will look at representations of the variables in terms of independent latent varables $l_1, \ldots, l_q$, $\varepsilon, \delta_1, \ldots, \delta_p$ as

$$\begin{aligned} y &= \sum_{k=1}^{m} b_k l_k + \varepsilon \\ x_j &= \sum_{k=1}^{q} a_{jk} l_k + \delta_j \qquad (j = 1, \ldots, p) \end{aligned} \quad , \qquad (5.1)$$

where $m \le q \le p$ and $\mathrm{Var}(\delta_1) = \ldots = \mathrm{Var}(\delta_p) = \tau$. Given the covariance structure of $y, x_1, \ldots, x_p$, a representation of the form (5.1) is always possible to find, and it is easy to characterize the minimal values of m and q and to give an expression for a possible choice of latent variables. Let $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_p > 0$ be the eigenvalues of Σ_{xx}. We will make the assumption

(N) The vector σ_{xy} has no component along the eigenspace corresponding to the minimal eigenvalue λ_p.

In the language of Næs and Helland (1993) this means that the smallest eigenvalue is not relevant for prediction. For data occurring in spectroscopy and in similar problems, this seems to be a weak assumption, at least it can easily be assumed as a simplification of the model structure. If we have collinear data where this can not be assumed, serious identification problems will occur.

Proposition 5.1.

a) It is always possible to find a representation (5.1) such that $m \le q \le p$, and then necessarily $\tau \le \lambda_p$. Under the condition (N) we can take $\tau = \lambda_p$ and then $q=p-r$, where r is the multiplicity of λ_p. This is the minimal value of q.

b) Assume (N) and $\tau = \lambda_p$. The minimal number m is equal to the number of distinct eigenvalues (larger than λ_p) whose eigenspaces have nonvanishing components in the direction σ_{xy}.

c) Assume (N) and assume $\tau = \lambda_p$ and m minimal. Let $\boldsymbol{x}=(x_1, \ldots, x_p)'$ and $\boldsymbol{d}= (\delta_1, \ldots, \delta_p)'$. Then $\boldsymbol{x-d}$ is stochastically independent of $\boldsymbol{d}$. Organize the eigenvectors $(\boldsymbol{e}_k; k = 1, \ldots, p)$ of Σ_{xx} such that $\boldsymbol{e}_k$ (k=1, ... ,m) have non-vanishing components in the direction σ_{xy} and such that $\boldsymbol{e}_{q+1}, \ldots, \boldsymbol{e}_p$ correspond to the eigenvalue λ_p. Then a possible choice of l_k is given by $\boldsymbol{e}_k'(\boldsymbol{x}-\boldsymbol{d})$, normalized. Furthermore, $\boldsymbol{a}_k = (a_{1k}, \ldots a_{pk})'$ will be along $\boldsymbol{e}_k$.

Proof.

With $m=q=p$ a representation of the form (5.1) can be found trivially by using a principal component representation in the last equation and taking all $\delta_j = 0$. Since (5.1) implies that $\Delta_{xx} = \Sigma_{xx} - \tau \boldsymbol{I}$ must be nonnegative definite, it follows that $\tau \le \lambda_p$. We can write $\boldsymbol{x}=\boldsymbol{z}+\boldsymbol{d}$ with $\boldsymbol{z}$ and $\boldsymbol{d}$ independent, $V(\boldsymbol{z})= \Delta_{xx}$ and $V(\boldsymbol{d})= \tau \boldsymbol{I}$. From this, c) and the last part of a) follow since the eigenspaces of Σ_{xx} and Δ_{xx} are equal (except for the eigenspace corresponding to λ_p). This, together with the results of Section 5 in Helland (1990) gives b).

◊

The important conclusion is that the minimal number m of uncorrelated latent variables connecting the x- and the y-variables is just the same as the number m in the definition of H_m in Section 2. The variables l_k [proportional to $\boldsymbol{e}_k'(\boldsymbol{x}-\boldsymbol{d})$] for k=1, ... ,m can

in a natural way be called the relevant factors, since these are the only latent variables from the $\boldsymbol{x}$-equation that also occur in the y-equation. In Helland (1990, 1992) and in Næs and Helland (1993) the corresponding variables $\boldsymbol{e}_k'\boldsymbol{x}$ are called relevant components.

The latent variables in (5.1) are chosen on the basis of principal components. As shown in Helland (1990), a completely equivalent representation can be given in terms of partial least squares scores and loadings, found via the algorithm

$$\begin{aligned}
&\boldsymbol{x}^{(0)} = \boldsymbol{x}, \quad y^{(0)} = y, \text{ and for } a=1,2,\ldots: \\
&\boldsymbol{w}_a = \mathrm{Cov}(\boldsymbol{x}^{(a-1)}, y^{(a-1)}), \qquad t_a = \boldsymbol{x}^{(a-1)'}\boldsymbol{w}_a, \\
&\boldsymbol{p}_a = \mathrm{Cov}(\boldsymbol{x}^{(a-1)}, t_a)/\mathrm{Var}(t_a), \\
&q_a = \mathrm{Cov}(y^{(a-1)}, t_a)/\mathrm{Var}(t_a), \\
&\boldsymbol{x}^{(a)} = \boldsymbol{x}^{(a-1)} - \boldsymbol{p}_a t_a, \qquad y^{(a)} = y^{(a-1)} - q_a t_a .
\end{aligned} \tag{5.2}$$

Proposition 5.2.

Assume (N) and assume $\tau = \lambda_p$ and m minimal. Then the representation (5.1) is equivalent with

$$\begin{aligned}
y &= \sum_{a=1}^{m} q_a t_a + \varepsilon \\
x_j &= \sum_{a=1}^{m} p_{ja} t_a + \varepsilon_j
\end{aligned} \tag{5.3}$$

in the following sense:

$$\begin{aligned}
&\sum_{a=1}^{m} q_a t_a = \sum_{k=1}^{m} b_k l_k , \\
&\sum_{a=1}^{m} p_{ja} t_a = \sum_{k=1}^{m} a_{jk} l_k + \delta_j^{(rel)}, \qquad \varepsilon_j = \sum_{k=m+1}^{q} b_{jk} l_k + \delta_j^{(irrel)}, \\
&\mathrm{span}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m) = \mathrm{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m) = \mathrm{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m) .
\end{aligned}$$

Here $\boldsymbol{d}^{(rel)} = (\delta_1^{(rel)}, \ldots, \delta_p^{(rel)})'$ is the projection of $\boldsymbol{d}$ upon the relevant space $\mathrm{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$, and $\boldsymbol{d}^{(irrel)} = (\delta_1^{(irrel)}, \ldots, \delta_p^{(irrel)})' = \boldsymbol{d} - \boldsymbol{d}^{(rel)}$.

Furthermore, $\boldsymbol{p}_a = (p_{1a}, \ldots, p_{pa})'$. The scores $t_1, \ldots, t_m$ are uncorrelated.

Proof.

This follows from Helland (1990, Teorem 4). As before, let $\boldsymbol{e}_1, \ldots \boldsymbol{e}_p$ be the eigenvectors of Σ_{xx}, with the relevant ones first, and note that

$$\sum_{k=1}^{m} \boldsymbol{a}_k l_k = \sum_{k=1}^{m} \boldsymbol{e}_k \boldsymbol{e}_k' (\boldsymbol{x} - \boldsymbol{d}) = \sum_{k=1}^{m} \boldsymbol{e}_k (\boldsymbol{e}_k' \boldsymbol{x}) - \boldsymbol{d}^{(rel)} .$$

◊

6. Estimation

Let the data be given by the nx(p+1) matrix $(\boldsymbol{X},\boldsymbol{y})$, whose rows are independent and are assumed to have the covariance structure described above. By the sample partial least square algorithm as described for instance in Frank and Friedman (1993) one finds estimates $\hat{\boldsymbol{p}}_1, \ldots, \hat{\boldsymbol{p}}_m, \hat{q}_1, \ldots, \hat{q}_m$ and estimated scores $\hat{\boldsymbol{t}}_1, \ldots, \hat{\boldsymbol{t}}_m$. Consistency is nearly immediate.

Proposition 6.1.

Assume that m is minimal in the sense described in Section 2 and that all variables have finite covariances. Then as $n \to \infty$ we have

$$\hat{\boldsymbol{p}}_a \to_{a.s.} \boldsymbol{p}_a, \quad \hat{q}_a \to_{a.s.} q_a \quad (a = 1, \ldots, m)$$
$$n^{-1} \boldsymbol{I}' \hat{\boldsymbol{T}}_m \to_{a.s.} \boldsymbol{0} \quad \text{and} \quad n^{-1} \hat{\boldsymbol{T}}_m' \hat{\boldsymbol{T}}_m \to_{a.s.} \mathrm{Cov}(t_1, \ldots, t_m),$$
$$\text{where } \hat{\boldsymbol{T}}_m = (\hat{\boldsymbol{t}}_1, \ldots, \hat{\boldsymbol{t}}_m)$$

Proof.

By formulae in Frank and Friedman (1993) and Helland (1988, 1990, 1992) the estimated loadings can found by finite algorithms starting from $\boldsymbol{S}_{xx} = n^{-1}\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{s}_{xy} = n^{-1}\boldsymbol{X}'\boldsymbol{y}$ and the theoretical loadings by the same algorithms starting from Σ_{xx} and σ_{xy}. Furthermore, in terms of the weight factors $\hat{\boldsymbol{W}}_m = (\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_m)$ found by $\hat{\boldsymbol{w}}_1 = \boldsymbol{s}_{xy}$ and $\hat{\boldsymbol{w}}_{a+1} = \boldsymbol{s}_{xy} - \boldsymbol{S}_{xx}\hat{\boldsymbol{W}}_a(\hat{\boldsymbol{W}}_a'\boldsymbol{S}_{xx}\hat{\boldsymbol{W}}_a)^{-1}\hat{\boldsymbol{W}}_a'\boldsymbol{s}_{xy}$ we have $\hat{\boldsymbol{T}}_m = \boldsymbol{X}\hat{\boldsymbol{W}}_m(\hat{\boldsymbol{P}}_m'\hat{\boldsymbol{W}}_m)^{-1}$ and hence $\hat{\boldsymbol{T}}_m'\hat{\boldsymbol{T}}_m = (\hat{\boldsymbol{W}}_m'\hat{\boldsymbol{P}}_m)^{-1}\hat{\boldsymbol{W}}_m'\boldsymbol{S}\hat{\boldsymbol{W}}_m(\hat{\boldsymbol{P}}_m'\hat{\boldsymbol{W}}_m)^{-1}$. The proposition then follows by an application of the law of large numbers and Slutskys theorem.

◊

By the same argument the vector of regression coeffisients estimated by an m steps partial least squares algorithm is almost surely consistent under the assumptions given above. One can show from this that the crossvalidation function defined by Frank and Friedman (1993) will have a minimum for large n when m terms are included in the prediction.

None of the above arguments require that the residual vector $\boldsymbol{d} = (\delta_1, \ldots \delta_p)'$ in (5.1) should be small in any sense. The only difficulty that arises when $\boldsymbol{d}$ is not small, is that correspondence between the representations (5.1) and (5.3) is a little more complicated. The collinearity that one sees in chemometrical data, can usually be traced back to the fact that p is large compared to n. A model of the form (5.1) with non-negligible $\boldsymbol{d}$ may well be used both to explain the structure of and to simulate such data. If the population model is of this form, we have a situation where there exist simple consistency results for partial least squares regression, but not similarly good results for ridge regression as shown by Corollary 3.1.

We have tried to show in this paper that the model with m relevant components appears to be important in many ways when discussing regression methods. An obvious, but fundamental question remains, however: To what degree is it natural to assume such a model when analyzing real data? As a partial answer to this, we have the following remarks, all quite straightforward, but important enough to be repeated:

1) The model under discussion is in fact a nested collection of models, indexed by m. When $m=p$ the model is not restricted at all (except of course that we impose the conditions of linearity and normality); in general the value of m can be chosen by crossvalidation.

2) The problem with collinearity is closely related to the fact that the full model has too many parameters. Hence any sensible scheme for reduction of the number of parameters may be useful. The asymptotic prediction error that resulted from such a model reduction was discussed in Section 2 above, and simple formulas like (2.4) seem to indicate that the present restriction is reasonable in many cases.

3) The way we have reduced the model from the full model here, corresponds to reducing the minimal number of orthogonal latent variables that can be used to describe the coupling between the x- and y-variables.

References

Björck, Å. (1967). Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT* **7**, 1-21.

Frank, I.E., and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (incl. discussion). *Technometrics* **35**, 109-148.

Helland, I.S. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics* **43**, 61-69.

Helland, I.S. (1988). On the structure of partial least squares regression. *Communications in Statistics-Simulation and Computation* **17**, 581-607.

Helland, I.S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics* **17**, 97-114.

Helland, I.S. (1992). Maximum likelihood regression on relevant components. *Journal of the Royal Statistical Society B* **54**, 637-647.

Helland, I.S., and Almøy, T. (1994). Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association* **89**, 583-591.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Lawley, D.N., and Maxwell, A.E. (1973). Regression and factor analysis. *Biometrika* **60**, 331-338.

Martens, H., and Næs, T. (1989). *Multivariate Calibration*, Wiley, New York

Næs, T., and Helland, I.S. (1993). Relevant components in regression. *Scandinavian Journal of Statistics* **20**, 239-250.

Stewart, G.W. (1969). On the continuity of the generalized inverse. *SIAM J. Appl. Math.* **17**, 33-45.

Stewart, G.W. (1973). *Introduction to Matrix Computation.* Academic Press, New York.

Appendix.

Proof of Theorem 2.3.

A straightforward expansion using $f(\beta,\boldsymbol{R}) = (\beta_R - \beta)' \Sigma_{xx}(\beta_R - \beta)$ with $\beta_R = (\boldsymbol{R}' \Sigma_{xx}\boldsymbol{R})^{-1}\boldsymbol{R}' \Sigma_{xx}\beta$ yields

$$\left|\mathrm{E}[(\hat{\beta} - \beta)' \Sigma_{xx}(\hat{\beta} - \beta)] - f(\beta,\boldsymbol{R})\right| \le 2\{\mathrm{E}[(\hat{\beta} - \beta_R)' \Sigma_{xx}(\hat{\beta} - \beta_R)]\}^{\frac{1}{2}} f(\beta,\boldsymbol{R})^{\frac{1}{2}} + \mathrm{E}[(\hat{\beta} - \beta_R)' \Sigma_{xx}(\hat{\beta} - \beta_R)]$$

From the polarization inequality we find

$$\mathrm{E}[(\hat{\beta} - \beta_R)' \Sigma_{xx}(\hat{\beta} - \beta_R)] \le 2\mathrm{E}[(\hat{\beta} - \hat{\beta}_R)' \Sigma_{xx}(\hat{\beta} - \hat{\beta}_R)] + 2\mathrm{E}[(\hat{\beta}_R - \beta_R)' \Sigma_{xx}(\hat{\beta}_R - \beta_R)], \quad \text{(A.1)}$$

where $\hat{\beta}_R = \boldsymbol{R}(\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{X}\boldsymbol{R})^{-1}\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{y}$. The last term here can be explicitly calculated (cf. formula (2) in Helland and Almøy (1994)) as

$$\mathrm{E}[(\hat{\beta}_R - \beta_R)' \Sigma_{xx}(\hat{\beta}_R - \beta_R)] = \sigma^2 m/(n - m - 1). \quad \text{(A.2)}$$

The first term on the righthand side of (A.1), where an estimate involving $\hat{\boldsymbol{R}}$ is compared to an estimate involving $\boldsymbol{R}$, can be bounded above by using results from numerical analysis (Björk, 1967; similar results are given by Steward, 1969, 1973). This requires a change in notation as follows:

$$\boldsymbol{A}=\boldsymbol{X}\boldsymbol{R}\,, \qquad \delta\boldsymbol{A}=\boldsymbol{X}(\hat{\boldsymbol{R}}-\boldsymbol{R}), \qquad \boldsymbol{b}=\boldsymbol{y}, \qquad \boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{b} = (\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{X}\boldsymbol{R})^{-1}\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{y},$$
$$\bar{\boldsymbol{x}} = (\boldsymbol{A} + \delta\boldsymbol{A})^{+}\boldsymbol{b} = (\hat{\boldsymbol{R}}' \boldsymbol{X}' \boldsymbol{X}\hat{\boldsymbol{R}})^{-1}\hat{\boldsymbol{R}}' \boldsymbol{X}' \boldsymbol{y}, \qquad \boldsymbol{r}=\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}, \qquad \delta\boldsymbol{r} = \boldsymbol{X}(\hat{\beta} - \hat{\beta}_R).$$

The QR-decomposition of $\boldsymbol{A}$ ($\boldsymbol{A}=\boldsymbol{Q}\mathcal{R}$) implies (all norms are Euclidean):

$$\left\|\mathcal{R}^{-1}\right\|^2 = \left\|(\mathcal{R}'\mathcal{R})^{-1}\right\| = \left\|(\boldsymbol{A}' \boldsymbol{A})^{-1}\right\| = \left\|(\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{X}\boldsymbol{R})^{-1}\right\| = \kappa(\boldsymbol{X}\boldsymbol{R})^2/\|\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{X}\boldsymbol{R}\|,$$

where $\kappa(\boldsymbol{XR})$ is the condition number for the *nxm* matrix $\boldsymbol{XR}$. The definition (7.5) and the inequality (7.7) of Björck (1967) then give

$$(\hat{\beta} - \hat{\beta}_R)' \boldsymbol{X}' \boldsymbol{X}(\hat{\beta} - \hat{\beta}_R) \leq 2\delta_n \|\boldsymbol{y}\|^2 \Big/ \left[1 - (\sqrt{2}+1)\sqrt{\delta_n}\right], \tag{A.3}$$

provided $(\sqrt{2}+1)\sqrt{\delta_n} < 1$, where $\delta_n = \kappa(\boldsymbol{XR})^2 \left\|(\hat{\boldsymbol{R}} - \boldsymbol{R})' \boldsymbol{X}' \boldsymbol{X}(\hat{\boldsymbol{R}} - \boldsymbol{R})\right\| \Big/ \|\boldsymbol{R}' \boldsymbol{X}' \boldsymbol{XR}\|$. Since $\boldsymbol{X}\hat{\beta}$ and $\boldsymbol{X}\hat{\beta}_R$ both are projections of $\boldsymbol{y}$, the upper bound $4\|\boldsymbol{y}\|^2$ is valid in all cases.

Now let $\boldsymbol{C}$ be the matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\Sigma_{xx}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. In general $\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{CXa} \leq \|\boldsymbol{C}\|\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{Xa}$, and therefore from (A.3)

$$(\hat{\beta} - \hat{\beta}_R)' \Sigma_{xx}(\hat{\beta} - \hat{\beta}_R) \leq 2\|\boldsymbol{C}\|\|\boldsymbol{y}\|^2 \min\{2, \delta_n \Big/ \left[1 - (\sqrt{2}+1)\sqrt{\delta_n}\right]_+\}. \tag{A.4}$$

The norm of $\boldsymbol{C}$ is the largest eigenvalue of $\boldsymbol{C}$. By simple manipulation of the eigenvalue equation, this is seen to be equal to the largest eigenvalue, hence the norm of $\Sigma_{xx}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Inserting (A.2) and (A.4) into (A.1) and then into the equation above (A.1) completes the proof.

Figure caption.

Fig. 1: Contribution to asymptotical prediction error [cf. equations (2.5)-(2.6)] from irrelevant component as a function of the corresponding eigenvalue. Relevant eigenvalues: $\lambda_1 = 1.0, \quad \lambda_2 = 2.0$; corresponding regression coefficients: $\eta_1 = 0.5, \quad \eta_2 = 0.3$. Residual variance: $\sigma^2 = 0.36$.

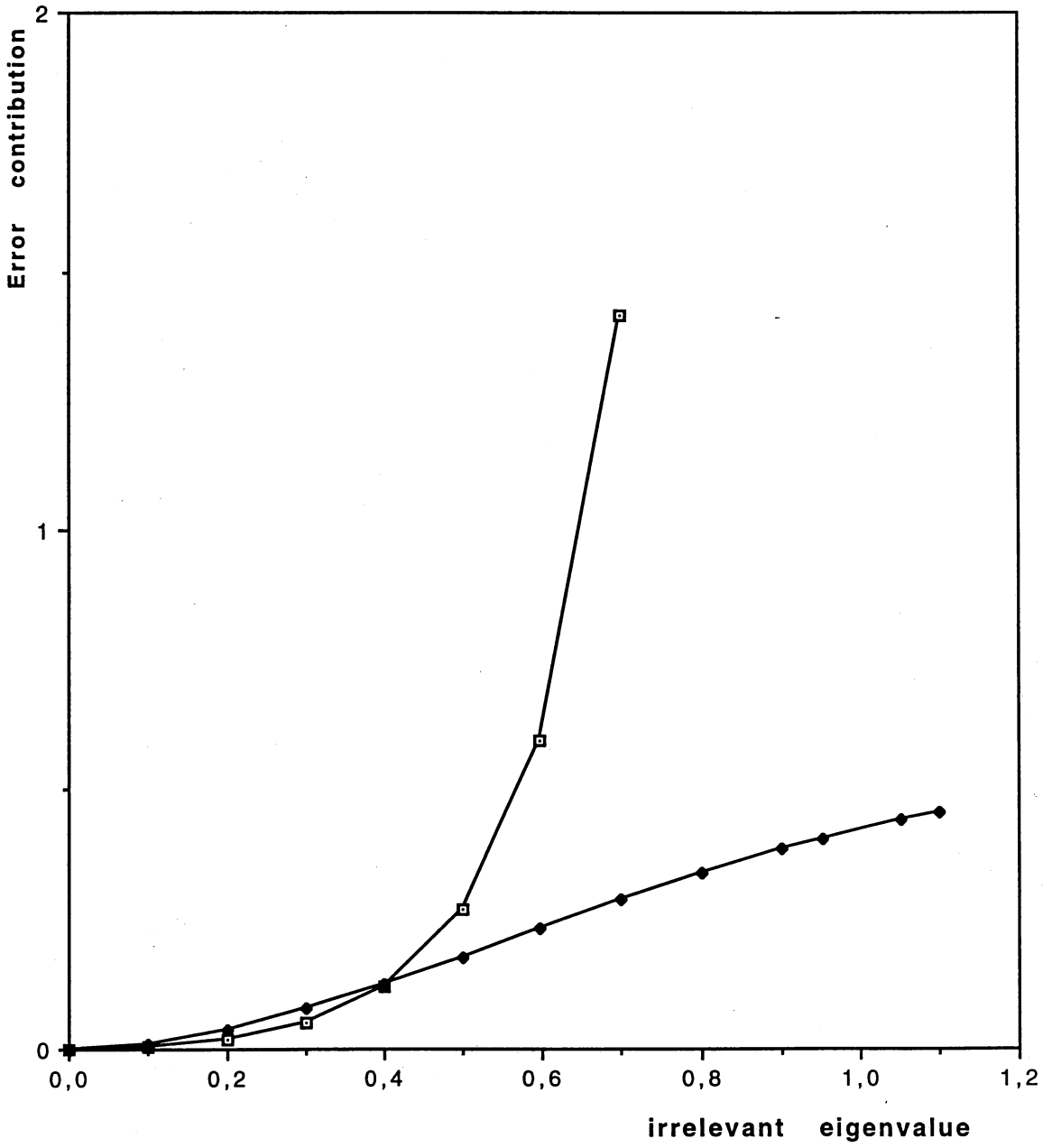
Error contribution
2
1
0
0,0
0,2
0,4
0,6
0,8
1,0
1,2
irrelevant eigenvalue

Fig. 1